Open Science intelligent wikis

Finn Årup Nielsen

Project description

Wikis have proliferated in knowledge-based organizations filling a need for quick easy online collaboration. However, their ability to manipulate numeric data is strikingly limited. To explore this aspect this project will construct **intelligent wikis**. As well as textual information such wikis can represent structured information, make queries, perform numerical computations, make numerical-based visualization and export data in a variety of formats.

Initially we will apply such wikis in **neuroscience**. This field provides sufficiently complex, dynamic and public data for exploring the utility of the approach. Our first aim is to develop a system that enables online collaborative meta-analysis of neuroscience data, while the goal is a general system for collaborative interaction with both numerical data and textural information.

The project will build extraction tools to populate databases as well as build ontologies to organize the data. The project will build on an already established framework for databasing of functional and molecular neuroimaging, but extend it with Web 2.0 services allowing researchers to upload result data from within image analysis programs, perform online meta-analyses and browse the spatial data in three dimensional visualizations. An API will be constructed enabling third party applications to download and manipulate the data. The focus is on **Open Science** where (paraphrasing Michael Nielsen) scientific knowledge, with its data and methods, is shared as early as possible in the discovery process.

Background

Wikis have become increasingly popular. Devised in the middle of the 1990s by Ward Cunningham (Leuf and Cunningham, 2001) wikis became generally known through the online encyclopedia Wikipedia initiated in January 2001: The multilingual Web-site is now among the top ten most popular Web sites world-wide. Many knowledge-based organizations have embraced the wiki-idea and setup a system on their intranet to let employees have a convenient method for sharing information that should have some persistence, e.g., information about research procedures in a university research unit or technical information for users of the IT system of an organization. Many organizations that need to share information with outsiders such as software co-developers also setup a wiki on the public Internet, and scientists share knowledge through Wikipedia **(Nielsen, 2007)**[1] as well as on a number of other specialized scientific wikis. Wiki functionality may be provided by MediaWiki, the software running Wikipedia, but a large number of other software packages also provide wiki functionality. Some of these, including MediaWiki, enable third-party extensions that provide extra specialized functionality on the wiki. **Semantic wikis** provide Semantic Web functionality such that, e.g., wikilinks on the page can be typed and advanced queries can be made.

[1] Self citations are in bold typeface.

The **Semantic Web**, that seeks to describe Web content in a machine readable form (Berners-Lee et al., 2001), can form large-scale data sets which may have great utility for general data mining: According to IBM their "Watson" question answering system, that in February 2011 beat human experts in American television *Jeopardy*, was partly based on knowledge available from the Semantic Web. Essentially a Web-based three elements (triplet) data structure with consistent Web-reachable identifiers (HTTP URIs) and open standard languages such as RDF, Notation3 and SPARQL the Semantic Web may form **Linked Data** when multiple data sets are combined. One such data set, the DBpedia, extracts data from the Wikipedias (Auer et al., 2007), and with Web services exposes the extracted data matched to ontologies, so third-party developers may use the enormous amount of structured information that resides in Wikipedia. *BBC* (Kobilarov et al., 2009), *New York Times* and Danish newspaper *Information* have utilized the data provided by DBpedia and made it available on their homepages. Many other Linked Data data sets exist that form the cloud of Linked Data, e.g., general databases such as Freebase, but also specialized life science databases such as PubMed and OMIM. The presently largest efforts in neuroscience are NeuroCommons and NeuroLex. With the Brede Wiki I have made initial efforts in generating data for the Semantic Web via the SKOS format **(Nielsen, 2009a)**.

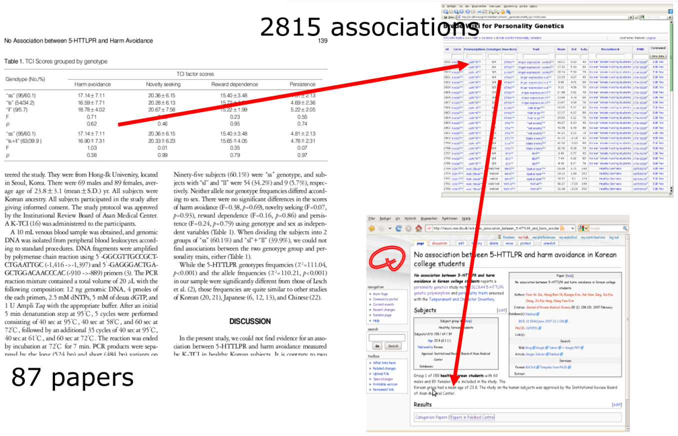
Neuroinformatics develops tools for analysis and visualization of neuroscience data. Most results of the neuroscience field is only reported in scientific papers—not uploaded to neuroinformatics databases—and database curators have difficulties in keeping up with the increasing amount of data being generated. I have constructed Brede Database based on XML and Matlab for results from neuroimaging studies and ontologies **(Nielsen, 2003)**. With information from 186 papers it is smaller than the BrainMap database and SumsDB, that each have information from around 2000 papers. However, even these large databases cannot keep up with the papers published in the neuroimaging field (Derrfuss and Mar, 2009), — and neuroimaging is just a small part of neuroscience. A recent *Science* article (Akil et al., 2011) reviewed the challenges and opportunities in neuroinformatics and made recommendations, e.g.:

> *1) Neuroscientists should, as much as is feasible, share their data in a form that is machine-accessible, such as through a Web-based database or some other structured form that benefits from increasingly powerful search tools. 2) Databases spanning a growing portion of the neuroscience realm need to be created, populated, and sustained. This effort needs adequate support from federal and other funding mechanisms. [...] 6) Investment needs to occur in interdisciplinary research to develop computational, machine-learning, and visualization methods for synthesizing across spatial and temporal information tiers.*

I have advocated for a more collaborative and wiki-oriented approach, **"neuroinformatics 2.0"**, to overcome the problem of database entry **(Nielsen, 2009b)**, and set up the MediaWiki-based Brede Wiki to record text and numerical information from neuroscience **(Nielsen, 2009a)**. Like a system such as AcaWiki the Brede Wiki allows for collaborative entry of bibliographic data, summary and comments on scientific papers, but the Brede Wiki also enables representing numerical data that can be extracted and added to standard database engines for efficient queries.

Online meta-analytic databases AlzGene (Bertram et al., 2007), SzGene, PDGene ALSGene and MSGene record data from genetic association studies on Alzheimer, schizophrenia, Parkinson diseases, ALS and Multiple Sclerosis respectively. The original data from the scientific studies are entered and presented in a **Web-based environment with meta-analytic results and visualizations**. However, these systems present little two-way interaction, and users of the database may apparently not manipulate the data, e.g., add extra data or directly download the data in a structured format.

I have built a prototype system—a **structured wiki**—that enables collaborative entry, interactive meta-analysis and visualization with export of data in several formats **(Nielsen, 2010)**, see Figure 1. The structured wiki is though confined to a very specific scientific area (personality genetics) and lacks the flexibility for other types of data in neuroscience and close-by areas. This project will seek to extend the system, so multiple types of data can be handled, more general computation can be made over the Web and data can be added and modified seamlessly



No Association between 5-HTTLPR and Harm Avoidance 139

Table 1. TCI Scores grouped by genotype

Genotype (No./%)	TCI factor scores			
	Harm avoidance	Novelty seeking	Reward dependence	Persistence
"ss" (95/60.1)	17.14±7.11	20.36±6.15	15.40±3.48	[illegible]±2.13
"ls" (54/34.2)	16.59±7.71	20.28±6.13	[illegible]	4.69±2.36
"ll" (9/5.7)	18.78±4.02	20.67±7.58	15.22±1.99	5.22±2.05
F	0.71	[illegible]	0.23	0.55
p	0.62	0.46	0.95	0.74
"ss" (95/60.1)	17.14±7.11	20.36±6.15	15.40±3.48	4.81±2.13
"ls+ll" (63/39.9)	16.90±7.31	20.33±6.23	15.65±4.05	4.76±2.31
F	1.03	0.01	0.35	0.07
p	0.38	0.99	0.79	0.97

teered the study. They were from Hong-Ik University, located in Seoul, Korea. There were 69 males and 89 females, average age of 23.8±3.1 (mean±S.D.) yr. All subjects were Korean ancestry. All subjects participated in the study after giving informed consent. The study protocol was approved by the Institutional Review Board of Asan Medical Center. A K-TCI (16) was administered to the participants.

A 10 mL venous blood sample was obtained, and genomic DNA was isolated from peripheral blood leukocytes according to standard procedures. DNA fragments were amplified by polymerase chain reaction using 5′-GGCGTTGCCGCTCTGAATTGC (-1,416 ->-1,397) and 5′-GAGGGACTGAGCTGGACAACCCAC (-910 ->-889) primers (3). The PCR reaction mixture contained a total volume of 20 μL with the following composition: 12 ng genomic DNA, 4 pmoles of the each primers, 2.5 mM dNTPs, 5 mM of deaza dGTP, and 1 U Ampli *Taq* with the appropriate buffer. After an initial 5 min denaturation step at 95℃, 5 cycles were performed consisting of 40 sec at 95℃, 40 sec at 58℃, and 60 sec at 72℃, followed by an additional 35 cycles of 40 sec at 95℃, 40 sec at 61℃, and 60 sec at 72℃. The reaction was ended by incubation at 72℃ for 7 min. PCR products were separated by the long (524 bp) and short (484 bp) variants on

Ninety-five subjects (60.1%) were "ss" genotype, and subjects with "sl" and "ll" were 54 (34.2%) and 9 (5.7%), respectively. Neither allele nor genotype frequencies differed according to sex. There were no significant differences in the scores of harm avoidance (F=0.38, p=0.69), novelty seeking (F=0.07, p=0.93), reward dependence (F=0.16, p=0.86) and persistence (F=0.24, p=0.79) using genotype and sex as independent variables (Table 1). When dividing the subjects into 2 groups of "ss" (60.1%) and "sl"+"ll" (39.9%), we could not find associations between the two genotype group and personality traits, either (Table 1).

While the 5-HTTLPR genotypes frequencies (χ^2=111.04, p<0.001) and the allele frequencies (χ^2=110.21, p<0.001) in our sample were significantly different from those of Lesch et al. (2), those frequencies are quite similar to other studies of Korean (20, 21), Japanese (6, 12, 13), and Chinese (22).

DISCUSSION

In the present study, we could not find evidence for an association between 5-HTTLPR and harm avoidance measured by K-TCI in healthy Korean subjects. It is contrary to two

http://neuro.imm.dtu.dk/wiki/No_association_between_5-HTTLPR_and_harm_avoidar

No association between 5-HTTLPR and harm avoidance in Korean college students

No association between 5-HTTLPR and harm avoidance in Korean college students reports a personality genetics study on the SLC6A4 5-HTTLPR genetic polymorphism and personality traits assessed with the Temperament and Character Inventory.

Subjects

Subject group 1 (help)
Healthy Korean students
Subjects/♂/♀: 158 / 69 / 89
Age: 23.8 ±3.1 (-)
Nationality: Korean
Approval: Institutional Review Board of Asan Medical Center
Databases:

Paper (help)
No association between 5-HTTLPR and harm avoidance in Korean college students
Authors: Yeon Ho Joo, Heung Bum Oh, Byungsu Kim, Suk Hoon Jung, Jun Kyu Chung, Jin Pyo Hong, Chang Yoon Kim
Citation: Journal of Korean Medical Science 22 (1): 138-131. 2007 February
Database(s): PubMed
DOI: 10.3346/jkms.2007.22.1.138
PMCID: 2693551
Link(s):
Search
Web: Bing Google Yahoo! — Google PDF
Article: Google Scholar PubMed
Services
Format: BibTeX Template from PMID
Extract:

Group 1 of 158 **healthy Korean students** with 69 males and 89 females were included in the study. The Korean group had a mean age of 23.8. The study on the human subjects was approved by the Institutional Review Board of Asan Medical Center.

Results

Categories: Papers | Papers in PubMed Central

Figure 1: Extraction of data from scientific papers to a highly structured wiki with built-in meta-analytic calculations as well as export to a more "standard" wiki (the Brede Wiki) **(Nielsen, 2010)**.

There are several systems that allow Internet users to share structured data and perform computations and visualizations online. Google has a several Web services that implement aspects of the proposal: Google Fusion Tables, Google DataWiki, Google Public Data Explorer and Google Spreadsheet. DataWiki and Public Data Explorer are in their early stage. Fusion Table can store tables, make queries, export and construct some common forms of visualizations, e.g., scatter plots (Gonzalez et al., 2010). Other online databases are, e.g., Freebase and Fluidinfo. The systems may have limited ability for computation and specialized plotting as well as lacking a revision control system known from wikis. The *Semantic MediaWiki* extension to MediaWiki can store structured data, make queries and visualizations (Krötzsch et al., 2006). However, the present version lacks some more advanced query capabilities, cannot perform numerical computations on its data or tailor visualizations. The standardized SPARQL Semantic Web framework allows for more advanced queries and present effort aims at integrating it into Semantic MediaWiki. Other related work are the off-line spreadsheets pyspread and CoreCalc implementations (Sestoft, 2006) as well as a patent on "collaborative analysis of information" (Stefik and Bobrow, 2010) which builds on the so-called *Analysis of Competing Hypotheses* program and methodology used for evidence weighting in the intelligence agencies community. Apart from my effort in **(Nielsen, 2009a; Nielsen, 2010)** very few have apparently explored how wikis can be combined with numerical computations.

Research plan

1) A wiki-like system of extensible databasing and computation.

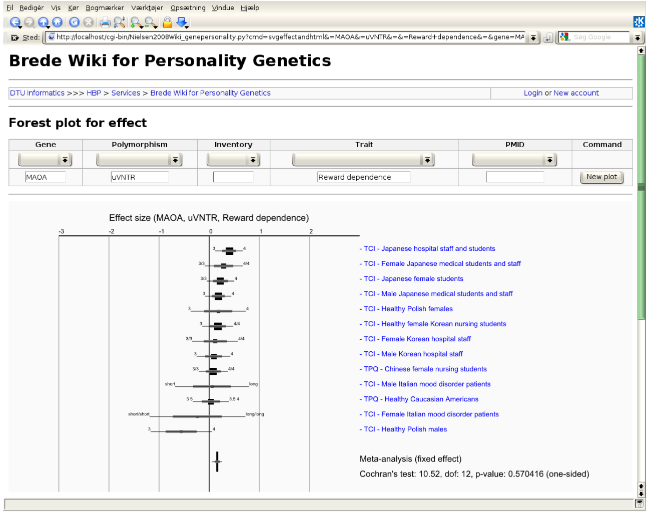
The wiki should be able to represent textual and numeric data in a table-like as well as a more free-form structure. Initially the wiki should be able to represent neuroinformatics data such as stereotaxic coordinates, neuroimaging result data, and genetic association data and perform online meta-analysis on this data. I envision three ways of building such a system:

1. Combining Semantic MediaWiki with SPARQL and further constructed MediaWiki extensions, e.g., with meta-analytic plotting extensions.

2. Representing data in simple formats, such as templates and comma-separated values, in standard MediaWiki and develop off-wiki online Web service the perform the computation.

3. Building a dedicated extensible wiki.

Initially we will implement standard meta-analytic statistics algorithm (Hedges and Olkin, 1985; Hartung et al., 2008) that works on standardized mean differences and logarithmic odds ratios as well as kernel density-based neuroimaging-specific meta-analysis **(Nielsen and Hansen, 2002)**. Later we will extend the system with more general computational capabilities with execution of collaboratively written code running in "sandboxes" either client side (browser) or server side (Web server). Researchers from the *Institute of Psychiatry* in London have performed large neuroimaging meta-analyses on major depressive and bipolar disorders (Kempton et al., 2008; Kempton et al., 2011). In a recent collaboration we have added part of their data to our MediaWiki-based wiki and constructed an off-wiki Web service with some of the standard meta-analytic statistics capability including forest and funnel plots, thus showing the feasibility of the approach. We will maintain and extend the data in the wiki so we have a backbone of interesting data for the further development of the system.


Fil Redigér Vis Kør Bogmærker Værktøjer Opsætning Vindue Hjælp
Sted: http://localhost/cgi-bin/Nielsen2008Wiki_genepersonality.py?cmd=svgeffectandhtml&=MAOA&=uVNTR&=&=Reward+dependence&=&gene=MA
Søg Google

Brede Wiki for Personality Genetics

DTU Informatics >>> HBP > Services > Brede Wiki for Personality Genetics | Login or New account

Forest plot for effect

Gene	Polymorphism	Inventory	Trait	PMID	Command
MAOA	uVNTR		Reward dependence		New plot

Effect size (MAOA, uVNTR, Reward dependence)

-3
-2
-1
0
1
2
3 4
3/3 4/4
3/3 4
3 4
3 4
3 4/4
3/3 4/4
3 4
3/3 4/4
short long
3 5 3.5 4
short/short long/long
3 4

- TCI - Japanese hospital staff and students
- TCI - Female Japanese medical students and staff
- TCI - Japanese female students
- TCI - Male Japanese medical students and staff
- TCI - Healthy Polish females
- TCI - Healthy female Korean nursing students
- TCI - Female Korean hospital staff
- TCI - Male Korean hospital staff
- TPQ - Chinese female nursing students
- TCI - Male Italian mood disorder patients
- TPQ - Healthy Caucasian Americans
- TCI - Female Italian mood disorder patients
- TCI - Healthy Polish males

Meta-analysis (fixed effect)

Cochran's test: 10.52, dof: 12, p-value: 0.570416 (one-sided)

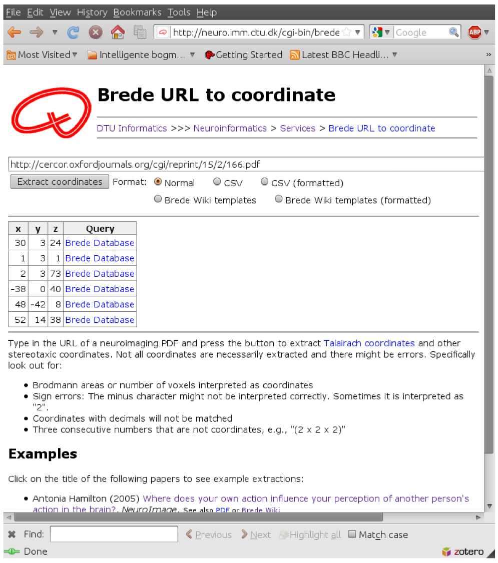
Figure 2: Online meta-analytic interactive plot with hyperlinks and results from personality genetics studies.

The final system should be able to perform Web-based mass meta-analysis over a diverse set of variables with collaboratively entered data in neuroscience and related fields, in a style similar to our previous off-line mass meta-analyses **(Nielsen, 2005; Nielsen et al., 2005)**. Specifically for the Institute of Psychiatry data we will expand it to include data from other neuropsychiatric disorders, so the mass meta-analysis can be performed across multiple brain regions and multiple disorders.

We will construct a Web service enabling Internet users with personal genomics genotyping to upload their raw results files (which, e.g., are available from the *23andme* personal genomics company) and compare it against the meta-analytic results of the our database.

2) Tools to extract data from scientific articles for populating databases.

This part will be extended from a simple Web service that with regular expression extracts brain coordinates from neuroscience papers represented in HTML or PDF. Recently, Tal Yarkoni has shown the feasibility of large-scale extraction in connection with his NeuroSynth web service for automated neuroinformatics databasing of neuroimaging studies (Yarkoni et al., 2011). Both his and our present systems only extract the three-dimensional coordinates and ignores the extra information available such as neuroanatomy and statistical values. We will extend it so more general structured data can be extracted from scientific articles.



Brede URL to coordinate

DTU Informatics >>> Neuroinformatics > Services > Brede URL to coordinate

http://cercor.oxfordjournals.org/cgi/reprint/15/2/166.pdf

Extract coordinates Format: ● Normal ○ CSV ○ CSV (formatted)
○ Brede Wiki templates ○ Brede Wiki templates (formatted)

x	y	z	Query
30	3	24	Brede Database
1	3	1	Brede Database
2	3	73	Brede Database
-38	0	40	Brede Database
48	-42	8	Brede Database
52	14	38	Brede Database

Type in the URL of a neuroimaging PDF and press the button to extract Talairach coordinates and other stereotaxic coordinates. Not all coordinates are necessarily extracted and there might be errors. Specifically look out for:

- Brodmann areas or number of voxels interpreted as coordinates
- Sign errors: The minus character might not be interpreted correctly. Sometimes it is interpreted as "2".
- Coordinates with decimals will not be matched
- Three consecutive numbers that are not coordinates, e.g., "(2 x 2 x 2)"

Examples

Click on the title of the following papers to see example extractions:

- Antonia Hamilton (2005) Where does your own action influence your perception of another person's action in the brain?. *NeuroImage*. See also PDF or Brede Wiki

Figure 3: Web service with the extraction of coordinates from a neuroimaging paper.

We will determine to which extent machine learning algorithms can be used to classify whether documents contain data that are relevant to extract, e.g., whether a specific paper contains neuroimaging data. Machine learning may also be applied to detect relevant context within the articles, see, e.g., (Cafarella et al., 2008) and references therein. We will use machine learning classifiers that we previously have built for text mining **(Hansen et al., 2000; Hansen et al., 2011)**.

We will explore if the Amazon Mechanical Turk can be used to populate scientific databases. We will furthermore develop extensions to popular neuroimaging analysis tools, so their users may upload results to our database working from a framework we previously have established **(Wilkowski et al., 2009)**. This part of the project should help populate our database.

3) Extension of the neuroscience databases to the Semantic Web and inclusion in *Linked Data*.

We will extend the data in the Brede Wiki and ensure it maps and integrates to other databases in the *Linked Data* cloud by extending it with Semantic Web mappings, establish persistent URIs for components, publish the data in triplet format and setup a SPARQL server allowing Internet users to query our database. We will seek to utilize standard components from Free and Open Source software, e.g., Python RDF libraries (Segaran et al., 2009), dedicated triplet data stores and integrate it with the wiki.

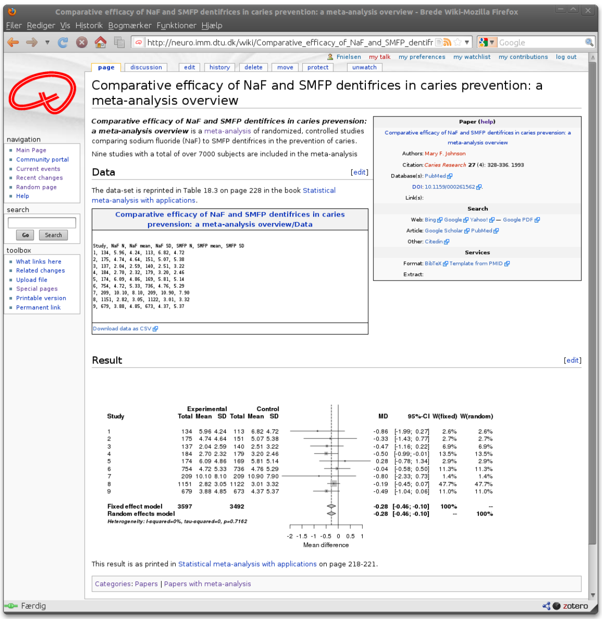
We will explore the suitability of cloud-based services such as Fluidinfo, Freebase, Google services, DataWiki, Public Data Explorer, Fusion Tables, for distributing our curated data. For example by replicating the data in these services.

Benefits

Quantitative statistical combinations of scientific research results are important for evaluating new science. An epidemiologist stated in *Science* in 1990 “[meta-analysis is] a boon for policy policy-makers who find themselves faced with a mountain of conflicting studies” (Mann, 1990). The **final system** should be a tool for more effective collaboration among research groups world-wide and an accelerated scientific knowledge accumulation and dissemination.

The project should result in tools that are general enough to not only be employed for scientific meta-analysis. As many organizations have embraced wiki a part of them may also find it useful to have “intelligent wikis” with a capability to process numerical data.

The project should be viewed as a part of the **Open Data** and **Open Science** movements where data and algorithms are made readily available on the Internet for researchers as well as the layman to manipulate, download and used in context not initially anticipated.


Comparative efficacy of NaF and SMFP dentifrices in caries prevention: a meta-analysis overview - Brede Wiki-Mozilla Firefox
Filer Rediger Vis Historik Bogmærker Funktioner Hjælp
http://neuro.imm.dtu.dk/wiki/Comparative_efficacy_of_NaF_and_SMFP_dentifr
Google
Fnielsen my talk my preferences my watchlist my contributions log out
page discussion edit history delete move protect unwatch
navigation
Main Page
Community portal
Current events
Recent changes
Random page
Help
search
Go Search
toolbox
What links here
Related changes
Upload file
Special pages
Printable version
Permanent link
Færdig
zotero

Comparative efficacy of NaF and SMFP dentifrices in caries prevention: a meta-analysis overview

Comparative efficacy of NaF and SMFP dentifrices in caries prevension: a meta-analysis overview is a meta-analysis of randomized, controlled studies comparing sodium fluoride (NaF) to SMFP dentifrices in the prevention of caries.

Nine studies with a total of over 7000 subjects are included in the meta-analysis

Paper (help)

Comparative efficacy of NaF and SMFP dentifrices in caries prevension: a meta-analysis overview

Authors: Mary F. Johnson

Citation: *Caries Research* **27** (4): 328-336. 1993

Database(s): PubMed

DOI: 10.1159/000261562.

Link(s):

Search

Web: Bing Google Yahoo! — Google PDF

Article: Google Scholar PubMed

Other: Citedin

Services

Format: BibTeX Template from PMID

Extract:

Data [edit]

The data-set is reprinted in Table 18.3 on page 228 in the book Statistical meta-analysis with applications.

Comparative efficacy of NaF and SMFP dentifrices in caries prevension: a meta-analysis overview/Data

```
Study, NaF N, NaF mean, NaF SD, SMFP N, SMFP mean, SMFP SD
1, 134, 5.96, 4.24, 113, 6.82, 4.72
2, 175, 4.74, 4.64, 151, 5.07, 5.38
3, 137, 2.04, 2.59, 140, 2.51, 3.22
4, 184, 2.70, 2.32, 179, 3.20, 2.46
5, 174, 6.09, 4.86, 169, 5.81, 5.14
6, 754, 4.72, 5.33, 736, 4.76, 5.29
7, 209, 10.10, 8.10, 209, 10.90, 7.90
8, 1151, 2.82, 3.05, 1122, 3.01, 3.32
9, 679, 3.88, 4.85, 673, 4.37, 5.37
```

Download data as CSV

Result [edit]

	Experimental			Control						
Study	Total	Mean	SD	Total	Mean	SD	MD	95%-CI	W(fixed)	W(random)
1	134	5.96	4.24	113	6.82	4.72	-0.86	[-1.99; 0.27]	2.6%	2.6%
2	175	4.74	4.64	151	5.07	5.38	-0.33	[-1.43; 0.77]	2.7%	2.7%
3	137	2.04	2.59	140	2.51	3.22	-0.47	[-1.16; 0.22]	6.9%	6.9%
4	184	2.70	2.32	179	3.20	2.46	-0.50	[-0.99; -0.01]	13.5%	13.5%
5	174	6.09	4.86	169	5.81	5.14	0.28	[-0.78; 1.34]	2.9%	2.9%
6	754	4.72	5.33	736	4.76	5.29	-0.04	[-0.58; 0.50]	11.3%	11.3%
7	209	10.10	8.10	209	10.90	7.90	-0.80	[-2.33; 0.73]	1.4%	1.4%
8	1151	2.82	3.05	1122	3.01	3.32	-0.19	[-0.45; 0.07]	47.7%	47.7%
9	679	3.88	4.85	673	4.37	5.37	-0.49	[-1.04; 0.06]	11.0%	11.0%
Fixed effect model	**3597**			**3492**			**-0.28**	**[-0.46; -0.10]**	**100%**	--
Random effects model							**-0.28**	**[-0.46; -0.10]**	--	**100%**

Heterogeneity: I-squared=0%, tau-squared=0, p=0.7162

This result is as printed in Statistical meta-analysis with applications on page 218-221.

Categories: Papers | Papers with meta-analysis

Figure 4: Example from Brede Wiki with online meta-analytic results from a standard data set taken from (Hartung et al., 2008). Here the data is represented in a simple comma-separated values page on the wiki, the data downloaded and analyzed off-line in the *R* programming language and the resulting plot added to the relevant wikipage. The final system should enable direct analysis on the system.

Plan for publication

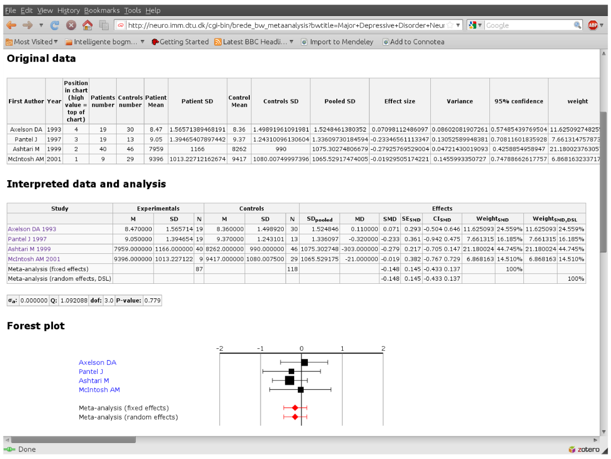
The project will result in a open online system with the primary part available from the Brede Wiki. We plan to publish three major articles related to the three subprojects, but we also expect other articles pertaining to details of the projects. During the project we will update a blog regularly. Data will be release under the Open Database License.

Facilities and environment

The project will be hosted at Section for Cognitive Systems, Department of Information and Mathematical Modelling, Technical University of Denmark. The department has **excellent computer facilities** with a Linux cluster and several high-memory servers for large-scale data mining. Further large-scale computing facilities exist at university level. I have several years experience with managing a Linux computer cluster as well as ordinary **system administration experience** on different Linux computers.

The section has built and operates several computers that provide different Web services on the public Internet, e.g., I manage services that extract information from scientific articles and PubMed formating the result in various ways. Also I have **experience in developing and maintaining**

neuroinformatics information retrieval Web-services that searches the special result data of neuroimaging either from the Web browser or from within an image analysis program. I have been operating a structured wiki for two and a half years.


File Edit View History Bookmarks Tools Help
http://neuro.imm.dtu.dk/cgi-bin/brede_bw_metaanalysis?bwtitle=Major+Depressive+Disorder+Neu
Google
Most Visited
Intelligente bogm...
Getting Started
Latest BBC Headli...
Import to Mendeley
Add to Connotea
-2
-1
0
1
2
Axelson DA
Pantel J
Ashtari M
McIntosh AM
Meta-analysis (fixed effects)
Meta-analysis (random effects)
Done
zotero

Original data

First Author	Year	Position in chart (high value = top of chart)	Patients number	Controls number	Patient Mean	Patient SD	Control Mean	Controls SD	Pooled SD	Effect size	Variance	95% confidence	weight
Axelson DA	1993	4	19	30	8.47	1.56571389468191	8.36	1.49891961091981	1.5248461380352	0.07098112486097	0.08602081907261	0.57485439769504	11.62509274825
Pantel J	1997	3	19	13	9.05	1.39465407897442	9.37	1.24310096130604	1.33609730184594	-0.23346561113347	0.13052589948381	0.70811601835928	7.661314757873
Ashtari M	1999	2	40	46	7959	1166	8262	990	1075.30274806679	-0.27925769529004	0.04721430019093	0.4258854958947	21.18002376305
McIntosh AM	2001	1	9	29	9396	1013.22712162674	9417	1080.00749997396	1065.52917474005	-0.01929505174221	0.1455993350727	0.74788662617757	6.868163233717

Interpreted data and analysis

Study	Experimentals			Controls			Effects								
	M	SD	N	M	SD	N	SD_{pooled}	MD	SMD	SE_{SMD}	CI_{SMD}	$Weight_{SMD}$		$Weight_{SMD,DSL}$	
Axelson DA 1993	8.470000	1.565714	19	8.360000	1.498920	30	1.524846	0.110000	0.071	0.293	-0.504 0.646	11.625093	24.559%	11.625093	24.559%
Pantel J 1997	9.050000	1.394654	19	9.370000	1.243101	13	1.336097	-0.320000	-0.233	0.361	-0.942 0.475	7.661315	16.185%	7.661315	16.185%
Ashtari M 1999	7959.000000	1166.000000	40	8262.000000	990.000000	46	1075.302748	-303.000000	-0.279	0.217	-0.705 0.147	21.180024	44.745%	21.180024	44.745%
McIntosh AM 2001	9396.000000	1013.227122	9	9417.000000	1080.007500	29	1065.529175	-21.000000	-0.019	0.382	-0.767 0.729	6.868163	14.510%	6.868163	14.510%
Meta-analysis (fixed effects)			87			118			-0.148	0.145	-0.433 0.137		100%		
Meta-analysis (random effects, DSL)									-0.148	0.145	-0.433 0.137				100%

σ_a:	0.000000	Q:	1.092088	dof:	3.0	P-value:	0.779

Forest plot

Figure 5: Initial setup for the proposed system with a Web service that downloads data from the Brede Wiki, extracts numerical data, computes effect sizes and meta-analytic results with forest plot. Data from the large neuroimaging meta-analysis study by Matthew Kempton and co-workers at the Institute of Psychiatry, London (Kempton et al., 2011).

The section have many years of experience with developing software packages. I have built Web services in the Python programming language and teach a university course in this language. **Extensive Free and Open Source Software programming libraries exist** for the Python programming language that makes it the primary language for Web-based data mining due to its support for Internet and numerical and scientific computation. Also libraries exist that enable easy interaction with the Semantic Web, wikis (MediaWiki), Google APIs, Fluidinfo and modern NoSQL databases.

Courses at the section serve as a source of students making projects. We expect to be able attract further master level students in the "Open Science Intelligent wiki" research area.

The Section for Cognitive Systems has several staff members that work within brain research and has formal research collaborations with outside neuroscience researchers. I maintain **affiliation with the neuroscience group** Neurobiology Research Unit, Copenhagen University Hospital, Rigshopitalet, which has resulted in several publications in the last few years and collaborate within the multi-center Copenhagen-based Center for Integrated Molecular Brain Imaging (CIMBI). These groups may provide an environment for interaction. The Cognitive Systems section itself headed by Lars Kai Hansen has also a long history for attracting numerous master and PhD students creating a lively environment, — particularly in data mining and machine learning.

I have also years of experience with populating neuroinformatics databases myself as well as building the associated ontologies. Recently, collaboration with London-based Institute of Psychiatry has resulted in population of the wiki with one of the largest meta-analytic data sets within neuroimaging.

Work and time plan

In the first part of first year we will setup and update a new server for hosting a Web server with the wiki. We will then start a first implementation of a wiki for fixed meta-analytic modeling and plotting. In this initial effort we will focus on MediaWiki extensions for tabular data and standard meta-analytic statistics (publication: *WikiSym* meeting or the *Frontiers in Neuroinformatics* journal) working from the data of the large meta-analysis by the Institute of Psychiatry, London. We also begin the construct of a Web-based extraction service for neuroimaging data (publication: technical neuroscience journal). In the second half of 2012 we will start on developing the Linked Data representation of the content of the wiki, — both meta-analytic data as well as ontologies (publication: e.g., *Extended Semantic Web Conference* — ESWC). Also in the last part of 2012 we will begin to research in a wiki-like system for more general numerical processing, more specifically we plan to determine to which extent SPARQL and the Semantic MediaWiki can be used for this task (publication: e.g., *ESWC* or *WikiSym*).

In the second year we will research in more general extraction algorithms and populate the wiki with extracted material (publication: *Frontiers in Neuroinformatics* or more technical conference). We will later that year setup a service for Linked Data query of our data and other associated data as well as a server for the genetic association information in our databases (publication: neuroinformatics or bioinformatics journal). In the second year we furthermore plan to research into a wiki-like system with more general and flexible computational abilities (publication: e.g., *WikiSym* or *PLoS ONE*).

During the first part of the third year we will use the system in neuroscience applications with mass meta-analyses of the content of the wikis (publication: *NeuroImage* or similar journal). The last part of last year will be spend with writing up the remaining results.

References

Akil, H., Martone, M. E., and Van Essen, D. C. (2011). Challenges and opportunities in mining neuroscience data. *Science*, 331(6018):708–712. DOI: 10.1126/science.1199305.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). DBpedia: A nucleus for a web of open data. In *The Semantic Web*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735, Heidelberg/Berlin. Springer. Link. DOI: 10.1007/978-3-540-76298-0_52.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web. *Scientific American*, 284(5):28–37.

Bertram, L., McQueen, M. B., Mullin, K., Blacker, D., and Tanzi, R. E. (2007). Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nature Genetics*, 39(1):17–23. DOI: 10.1038/ng19340.

Cafarella, M. J., Halevy, A., Wang, D. Z., Wu, E., and Zhang, Y. (2008). WebTables: Exploring the power of tables on the Web. *Proceedings of the VLDB Endowment*, 1(1):538–549. DOI: 10.1145/1453856.1453916.

Derrfuss, J. and Mar, R. A. (2009). Lost in localization: The need for a universal coordinate database. *NeuroImage*, 48(1):1–7. DOI: 10.1016/j.neuroimage.2009.01.053.

Gonzalez, H., Halevy, A., Jensen, C. S., Langen, A., Madhavan, J., Shapley, R., and Shen, W. (2010). Google Fusion Tables: Data management, integration and collaboration in the cloud. In *Proceedings of the 1st ACM symposium on Cloud computing*, pages 175–180. New York, NY, USA, ACM. Link. DOI: 10.1145/1807128.1807158.

Hansen, L. K., Arvidsson, A., Nielsen, F. Å., Colleoni, E., and Etter, M. (2011). Good friends, bad news — affect and virality in Twitter. In *Future Information Technology*, volume 185 of *Communications in Computer and Information Science*, pages 34–43, Berlin. Springer. Link. DOI: 10.1007/978-3-642-22309-9_5.

Hansen, L. K., Sigurdsson, S., Kolenda, T., Nielsen, F. Å., Kjems, U., and Larsen, J. (2000). Modeling text with generalizable Gaussian mixtures. In *Proceedings of ICASSP'2000*, Piscataway, New Jersey. Institute of Electrical and Electronics Engineers. Link.

Hartung, J., Knapp, G., and Sinha, B. K. (2008). *Statistical Meta-Analysis with Applications*. Wiley Series in Probability and Statistics. Wiley, Hoboken, New Jersey.

Hedges, L. V. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press, Orlando, Florida.

Kempton, M. J., Geddes, J. R., Ettinger, U., Williams, S. C. R., and Grasby, P. M. (2008). Meta-analysis, database, and meta-regression of 98 structural imaging studies in bipolar disorder. *Archives of General Psychiatry*, 65(9):1017–1032. DOI: 10.1001/archpsyc.65.9.1017.

Kempton, M. J., Salvador, Z., Munafo, M. R., Geddes, J. R., Simmons, A., Frangou, S., and Williams, S. C. R. (2011). Structural neuroimaging studies in major depressive disorder: meta-analysis and comparison with bipolar disorder. *Archives of General Psychiatry*, 68(7):675–690. DOI: 10.1001/archgenpsychiatry.2011.60.

Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., Bizer, C., and Lee, R. (2009). Media meets semantic web — how the BBC uses DBpedia and Linked Data to make connections. In *The Semantic Web: Research and Applications*, volume 5554 of *Lecture Notes in Computer Science*, pages 723–737, Berlin/Heidelberg. Springer. Link. DOI: 10.1007/978-3-642-02121-3.

Krötzsch, M., Vrandečić, D., and Völkel, M. (2006). Semantic MediaWiki. In *The Semantic Web - ISWC 2006*, volume 4273 of *Lecture Notes in Computer Science*, pages 935–942, Berlin/Heidelberg. Springer. Link. DOI: 10.1007/11926078_68.

Leuf, B. and Cunningham, W. (2001). *The Wiki Way: Quick Collaboration on the Web*. Addison-Wesley, Boston.

Mann, C. (1990). Meta-analysis in the breech. *Science*, 249(4968):476–480. DOI: 0.1126/science.2382129.

Nielsen, F. Å. (2003). The Brede database: a small database for functional neuroimaging. *NeuroImage*, 19(2). Presented at the 9th International Conference on Functional Mapping of the Human Brain, June 19–22, 2003, New York, NY.

Nielsen, F. Å. (2005). Mass meta-analysis in Talairach space. In *Advances in Neural Information Processing Systems 17*, pages 985–992, Cambridge, MA. MIT Press. Link.

Nielsen, F. Å. (2007). Scientific citations in *Wikipedia*. *First Monday*, 12(8).

Nielsen, F. Å. (2009a). Brede Wiki: Neuroscience data structured in a wiki. In *Proceedings of the Fourth Workshop on Semantic Wikis — The Semantic Wiki Web*, volume 464 of *CEUR Workshop Proceedings*, pages 129–133, Aachen, Germany. RWTH Aachen University. Link.

Nielsen, F. Å. (2009b). Lost in localization: A solution with neuroinformatics 2.0. *NeuroImage*, 48(1):11–13. DOI: 10.1016/j.neuroimage.2009.05.073.

Nielsen, F. Å. (2010). A fielded wiki for personality genetics. In *Proceedings of the 6th International Symposium on Wikis and Open Collaboration*, New York, NY, USA. ACM. Link. DOI: 10.1145/1832772.1832795.

Nielsen, F. Å., Balslev, D., and Hansen, L. K. (2005). Mining the posterior cingulate: Segregation between memory and pain component. *NeuroImage*, 27(3):520–532. DOI: 10.1016/j.neuroimage.2005.04.034.

Nielsen, F. Å. and Hansen, L. K. (2002). Modeling of activation data in the BrainMap$^{\text{TM}}$ database: Detection of outliers. *Human Brain Mapping*, 15(3):146–156. DOI: 10.1002/hbm.10012.

Segaran, T., Evans, C., and Taylor, J. (2009). *Programming the Semantic Web*. O'Reilly.

Sestoft, P. (2006). A spreadsheet core implementation in C#. IT University Technical Report Series TR-2006-91, IT University of Copenhagen, Copenhagen, Denmark. Link.

Stefik, M. J. and Bobrow, D. G. (2010). Method and system for the collaborative analysis of information. United States Patent 7,761,409. Link.

Wilkowski, B., Szewczyk, M., Rasmussen, P. M., Hansen, L. K., and Nielsen, F. Å. (2009). Coordinate-based meta-analytic search for the SPM neuroimaging pipeline: The BredeQuery plugin for SPM5. In *Proceedings of the Second International Conference on Health Informatics*, pages 11–17, Setubal, Portugal. INSTICC Press. Link.

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., and Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, 8(8):665–670. DOI: 10.1038/NMETH.1635.